

**OPTIMIZING FINE-TUNING METHODS FOR BASE LARGE LANGUAGE
MODELS: A CASE STUDY AT SWISS GERMAN UNIVERSITY**

By

Theodore Natanael Damanik
12002002

BACHELOR'S DEGREE
in

INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

Revision after the Thesis Defense on 1 July 2024

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.



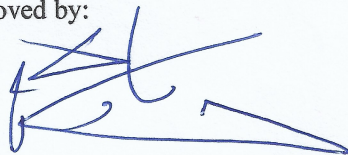
Theodore Natanael Damanik

Student

12/7/2024

Date

Approved by:

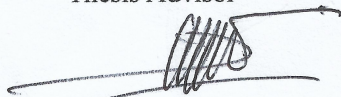


Dr. Maulahikmah Galinium, S.Kom, M.Sc

Thesis Advisor

12.7.2024

Date

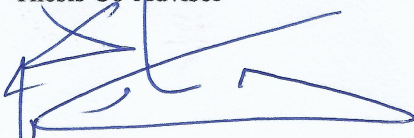


Assoc. Prof. Ir. James Purnama, S.Kom, M.Kom, M.Sc.

Thesis Co-Advisor

10/7/2024

Date



Dr. Maulahikmah Galinium, S.Kom, M.Sc

Dean

12.7.2024

Date

Theodore Natanael Damanik

ABSTRACT

OPTIMIZING FINE-TUNING METHODS FOR BASE LARGE LANGUAGE MODELS: A CASE STUDY AT SWISS GERMAN UNIVERSITY

By

Theodore Natanael Damanik

Dr. Maulahikmah Galinium, S.Kom, M.Sc, Advisor

Assoc. Prof. Ir. James Purnama, S.Kom, M.Kom, M.Sc., Co-Advisor

SWISS GERMAN UNIVERSITY

Chatbot and Large Language Models (LLM) are in growing trend recently. These chatbots can be used and customized for personal or commercial use, in this case for Swiss German University's customer acquisition team. To customize an LLM, one of the methods is to do fine-tuning. This thesis research is done to find the most compatible existing base LLM when fine-tuned with SGU's dataset which contains questions and answers pair about SGU. One of the problems of fine-tuning is that it requires a lot of computational resources. One of the solutions to this is to use a parameter efficient fine-tuning method named Low Rank Adaptation (LoRA). LoRA has multiple parameters that needs to be configured for fine-tuning efficiently. These parameter configurations are not fixed and varies depending on many variables. LoRA optimization is needed to find out the most optimal LoRA configurations for each base LLM. To do this, various LoRA configurations are created and used for fine-tuning the base LLM. The base LLMs used for this thesis research are Llama2 7B Chat by Meta and Mistral 7B Instruct by Mistral AI. Both base LLMs are tested with various LoRA configurations. Once the optimal configuration is found for both models, the two models are compared using ROUGE, BLEU, and METEOR. The fine-tuned Mistral 7B Instruct is able to achieve a higher score in all three metrics than the fine-tuned Llama2 7B Chat.

Keywords: Fine-tuning, LLM, Llama, Mistral, LoRA, Chatbot

© Copyright 2024
by Theodore Natanael Damanik
All rights reserved

DEDICATION

I dedicate this work for my family who always support me in anything.

ACKNOWLEDGEMENTS

I would like to first thank God for all the guidance during happy times and hard times. Especially during the process of this thesis research where I have encountered a lot of difficulties. I couldn't have done it without you, literally.

Second, I would like thank my family, especially my parents, that have been a really awesome guide throughout my life. My parents and siblings are very supporting in every way, even though sometimes I would get too hard headed to see that but at the end of the day they are always there for me when I need them. Thank you very much for all the support you have given to me and don't worry, I plan on giving even more support than you have given me. Not that it's a competition or anything.

Third, I would like to thank my advisor, Pak Maula, and my co-advisor, Pak James Purnama for providing all the support and consultation from the very first day of semester 1 in SGU until I do my thesis research in semester 8. I would like to also thank all the other lecturers in SGU that have taught me everything that I know. I have found my passion after studying in the classes you teach. Your lectures have become a foundation of everything I know about IT and I will never forget it. I would like to thank Kak Rana and Kak Aril for giving me an idea to do this thesis topic. This thesis topic is very interesting to me and I enjoyed the process of learning something new.

Lastly, I would like to thank my friends at SGU who have been beside me from semester 1, when covid hits and we could only see each other through our cameras and now we are all in semester 8. How the time flies. Thank you for all the memories we have created in Indonesia and in Germany too. Thank you for all the tutors and support throughout our studies. I wish you all good luck in the future.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR	2
ABSTRACT	3
DEDICATION	5
ACKNOWLEDGEMENTS	6
LIST OF FIGURES	9
LIST OF TABLES.....	10
CHAPTER 1 – INTRODUCTION.....	11
1.1 Background.....	11
1.2 Research Problems.....	12
1.3 Research Objectives	12
1.4 Significance of Study	12
1.5 Research Questions	13
1.6 Hypothesis	13
1.7 Research Scope.....	13
1.8 Research Limitation	13
CHAPTER 2 - LITERATURE REVIEW.....	15
2.1. Theoretical Perspectives.....	15
2.1.1. Natural Language Processing (NLP).....	15
2.1.2. Large Language Model (LLM)	15
2.1.3. ChatGPT	16
2.1.4. Flow-based.....	16
2.1.5. Quantization.....	17
2.1.6. Fine-Tuning.....	18
2.1.7. Evaluation metrics.....	21
2.1.7.1. F1 Score.....	21
2.1.7.2. BLEU	22
2.1.7.3. METEOR	22
2.1.7.4. ROUGE	23
2.2. Existing Base LLMs.....	24
2.2.1. GPT-4	25
2.2.2. GPT-3.5	25
2.2.3. Llama2	25

2.2.4.	FALCON	26
2.2.5.	Mistral 7B	26
2.3.	Related Works	27
CHAPTER 3 – RESEARCH METHODS		30
3.1.	Research overview	30
3.2.	Interview and data gathering	31
3.3.	Data analysis and dataset creation	31
3.4.	Fine-tuning configuration development.....	31
3.5.	Fine-tuning Execution.....	31
3.6.	Evaluation and analysis.....	33
CHAPTER 4 – RESULTS AND DISCUSSIONS		34
4.1.	Dataset.....	34
4.2.	Fine tuning configurations.....	34
4.2.1.	Data pre-processing	34
4.2.2.	Model settings	36
4.3.	Low Rank Adaptation Configuration Optimization	41
4.3.1.	Llama2 7B Chat Results	42
4.3.2.	Mistral 7B Instruct Results	43
4.4.	Llama2 7B Chat and Mistral 7B Instruct Question Answering Evaluation.....	45
4.5.	Proof of Concept.....	53
4.6.	Mixed language test	64
CHAPTER 5 – CONCLUSIONS AND RECCOMENDATIONS		67
5.1.	Conclusions	67
5.2.	Recommendations.....	68

LIST OF FIGURES

Figure 1. Current SGU's business process for customer acquisition.....	13
Figure 2. Low-Rank Adaptation.....	19
Figure 3. Research overview.....	30
Figure 4. Architecture diagram	32
Figure 5. Meta Llama 2 chat prompt format.....	35
Figure 6. Dataset formatting to follow Llama2 chat prompt format	35
Figure 7. Convert the dataset to huggingface dataset format.....	36
Figure 8. Base model name, new model name, and quantization settings	36
Figure 9. Loading the base model	37
Figure 10. Loading the tokenizer	37
Figure 11. LoRA config.....	37
Figure 12. Fine tuning training arguments.....	38
Figure 13. Creating the trainer	40
Figure 14. Training the model.....	41
Figure 15. Llama2 7B Chat LoRA configurations	42
Figure 16. Mistral 7B Instruct LoRA configurations	43
Figure 17. Shuffle the dataset for evaluation	46
Figure 18. Obtaining the answers from the model	46
Figure 19. Clean the generated answers	47
Figure 20. Load the metrics and evaluate the answers	47
Figure 21. All metrics and both model comparison	48
Figure 22. METEOR values of both models.....	49
Figure 23. ROUGE values of both models	49
Figure 24. Comparison of both models	50
Figure 25. BLEU values of both models	50
Figure 26. Code to generate answers from the models.....	53

LIST OF TABLES

Table 1. Other PEFT methods.....	20
Table 2. Table of popular base LLMs	24
Table 3. Strengths and weaknesses of related works.....	28
Table 4. Llama2 7B Chat LoRA Optimization Result	42
Table 5. Mistral 7B Instruct LoRA Optimization Result.....	44
Table 6. Llama2 7B Chat testing results with F1 Score	51
Table 7. Mistral 7B Instruct testing results with F1 Score	52
Table 8. Best LoRA Configurations for both fine-tuned models.....	67